ML 1.1 : VARIABLES, RANGE, POPULATION DISTRIBUTION, SAMPLE DISTRIBUTION (V) #600

PRESENTED BY:

DEEPTHI M

GITHUB LINK: HTTPS://GITHUB.COM/DEEPTHI1107

BATCH NUMBER: 02

SERIAL NUMBER: 198

Types of variables:

Independent Variable

Dependent Variable

y=f(x)
Where,
x= independent variable
y= dependent variable

Lets consider the following examples:

1. As temperature increases sales of ice-cream increases.
Here, sales of ice-cream is dependent on temperature.
Hence, y=f(x) in this formula:
y is sales of ice-cream.
x is temperature.

2. As working experience increases salary increases.
Here, salary is dependent on working experience.
Hence, y=f(x) in this formula:
y is salary.
x is working experience.

Range:

- Range of values or variables vary while dealing with machine learning algorithms as it includes dealing with different datasets.
- Size of datasets vary as it includes different objectives and business problems.

X1	X2	X3	X4

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10

When dealing with range of values, two main problems arises:

- What happens when we have wide range of values in different columns?
- What happens when we have wide range of columns?

X1	X2	X3	X4	X5
18389891038	78	5	789	77272
61786876276	67	78	9010	83892
1289833	8	890	83922	898918983
73981723873	90	73878	83	387828

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

When we have wide set of values in columns we treat this in two ways broadly:

- Standardization
- Normalization

Formula for Standardization:

(X – mean) / (std)

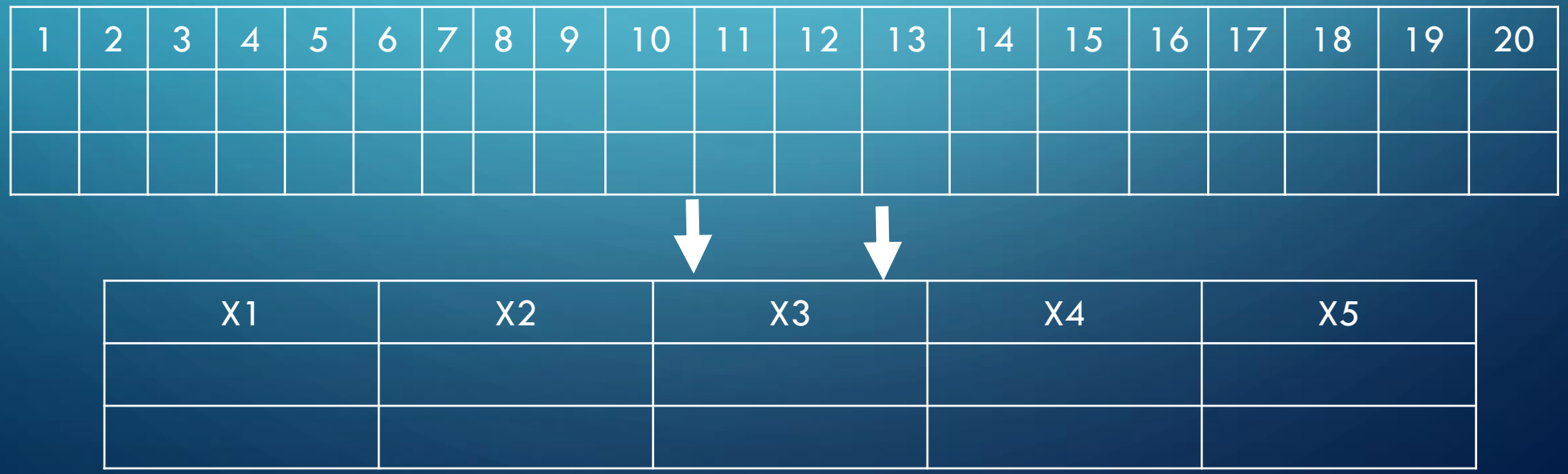
Formula for Normalization:

(X- X(min)) / (X(max) – X(min))

Here, X is input variable.

When we have wide range of columns we treat this by the following ways broadly:

- Principal Component Analysis or PCA
- Singular Value Decomposition or SVD
- Linear Discriminant Analysis or LDA



Population Distribution and Sample Distribution

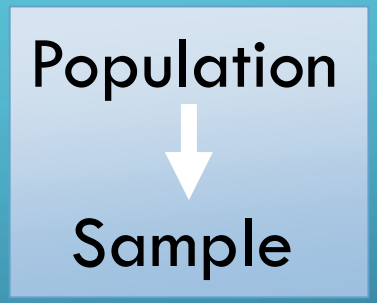
- The population is the whole set of values, or individuals, we are interested in.
- The sample is a subset of the population, and is the set of values we actually use in your estimation.

Lets understand with an example:

Business objective: to predict the height of an individual in India.

The dataset provided has heights of each individual in India.

Hence, dealing with entire population dataset ends up with problems like complexity, efficiency and accurate results. To avoid this we take sample from the dataset and apply algorithms for better performance.



Overview of sample distribution

- A sampling distribution is a statistic that is arrived out through repeated sampling from a larger population.
- It describes a range of possible outcomes that of a statistic, such as the mean or mode of some variable, as it truly exists a population.
- The majority of data analyzed by researchers are actually drawn from samples, and not populations.

Overview of population distribution

- Population distribution is a statistic arrived from considering the whole population.

Some of the measures.

Population Mean	Sample Mean
$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
N = number of items in the population	n = number of items in the sample

Some of the measures.

Sample Standard Deviation:

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{(N - 1)}}$$

Population Standard Deviation:

$$\sigma = \sqrt{\frac{\sum(x_i - \bar{x})^2}{N}}$$

Some of the measures.

$$\sigma^2 = \frac{\Sigma (x - \mu)^2}{N}$$ Population Variance

$$s^2 = \frac{\Sigma (x - \bar{x})^2}{n - 1}$$ Sample Variance

I hope now we have an overview on the topics:

- Variance
- Range
- Population distribution
- Sample distribution

Thank you I hope you have enjoyed!!!